Hans Journal of Data Mining 数据挖掘, 2016, 6(4), 148-157
Published Online October 2016 in Hans. http://www.hanspub.org/journal/hjdm
http://dx.doi.org/10.12677/hjdm.2016.64017

Analysis and Mining Based on the Online Recruitment Information

Lulu Zhang[1,2], Yumei Liao[1,2,3], Li He[1], Xiaomin Guo[1], Qian Liu[1]

[1]Department of Mathematics and Computer Science, Guizhou Education University, Guiyang Guizhou
[2]Internet + Innovation and Entrepreneurship Center of Guizhou Education University, Guiyang Guizhou
[3]Industrial Internet of Things Engineering Research Center of the Higher Education Institutions of Guizhou Province, Guizhou Education University, Guiyang Guizhou
Email: 1178334310@qq.com, liaoyumei-1999@163.com, 1561810892@qq.com, 971290510@qq.com, 1030550826@qq.com

Received: Sep. 28th, 2016; accepted: Oct. 15th, 2016; published: Oct. 18th, 2016

Abstract

With the development of Internet technology, online recruitment has become a communication platform for recruiters and applicants. In order to understand the dynamic information of society and the related industries, using the network text information and data mining technology to research the online recruitment information, we can get the social demand for talent and talent demand trends, which is of great significance to the colleges and universities and the graduates preparing for employment.

Keywords

Online Recruitment, Numerical Value, ARIMA Model, Grey Model

基于网络招聘信息的分析与挖掘

张路路[1,2]，廖玉梅[1,2,3]，何　丽[1]，郭小敏[1]，刘　倩[1]

[1]贵州师范学院数学与计算机科学学院，贵州　贵阳
[2]贵州师范学院大学生互联网+创新创业训练中心，贵州　贵阳
[3]贵州师范学院贵州省高校工业物联网工程技术研究中心，贵州　贵阳
Email: 1178334310@qq.com, liaoyumei-1999@163.com, 1561810892@qq.com, 971290510@qq.com, 1030550826@qq.com

文章引用: 张路路, 廖玉梅, 何丽, 郭小敏, 刘倩. 基于网络招聘信息的分析与挖掘[J]. 数据挖掘, 2016, 6(4): 148-157.
http://dx.doi.org/10.12677/hjdm.2016.64017

收稿日期：2016年9月28日；录用日期：2016年10月15日；发布日期：2016年10月18日

摘　要

随着网络信息技术的发展，网络招聘早已成为招聘者和应聘者交流的一大平台。为了了解社会和相关行业之间的动态信息，运用网络文本分析和数据挖掘技术对网络招聘信息进行研究，我们可以得到社会对人才的需求情况以及人才需求趋势，对高等院校以及应届毕业生的就业准备具有重要意义。

关键词

网络招聘，数值化，ARIMA模型，灰色模型

1. 引言

近几年来，传统招聘模式覆盖率低、效率差、成本高的弊端逐渐显现，但随着互联网对各个行业的渗透，各行各业开始倾向于采用互联网招聘的模式。其中智联招聘截止至 2016 年 6 月 6 日第四期总营业收入增长至 20.5%，足以可见，随着互联网时代的发展，不仅给网络招聘[1]带来了增长机会，而且招聘模式正在逐步上升。

本文利用招聘网站以及贵阳人力资源网，采用数据挖掘技术，并利用差分自回归移动平均模型、ARIMA 模型以及灰色预测法对人才市场的需求做出了预测，接着再对招聘职位的工作性质和内涵进行词云作图分析，可以得到职业类型与专业领域的分类，在此基础上分析其目前的人才需求情况，哪些是热门行业、职位等，并展望未来的人才需求走向。最后再使用与数据相关行业的职业进行分析，并预测未来大数据挖掘行业的发展趋势。

2. 分析方法与过程

由于网络招聘信息所涉及的数据信息较大，首先对文本[2]数据进行去重、去空，然后结合不同因素对文本数据进行图形的处理。本文的总体步骤及思路如下：

步骤一：首先对数据进行预处理，由于网络招聘信息所涉及的数据信息较大，因此可依据职业类型和行业领域的不同，可以将职业类型、行业领域进行简单的分类。利用 SPSS16.0 软件得到质量更高的数据，以便后续工作的分析。

步骤二：由步骤一处理后的数据分析目前的人才需求情况对职业类型和专业领域进行分类，建立 Excel 表格进行词频汇总，做简单的描述统计分析，接着利用 R 软件画出词云图，最后建立图形并进行分析，得出职业类型和专业领域的分类。

步骤三：基于 ARIMA 模型对销售岗位进行预测。

步骤四：通过对数据挖掘相关行业的分析，进行词频的分类，得出相关职位的需求情况，并用灰色预测法对其进行预测。

2.1. 数据数值化

由于网上发布的招聘信息不够严谨规范，因此存在许多不合理的数据，因此必须对实际数据进行数据数值化处理。本文通过词频的处理作出词云图，将数据进行数值化。

2.2. 职业类型与专业领域

对于招聘信息，本文首先对工作性质、工作内涵、职业类型以及专业领域进行了分析理解，并对收集中的数据进行了初步解读，并利用关键词的词频将每个职位样本进行数值化处理。明确分析行业领域、职业类型需求类型，借助 SPSS 读取 Excel 中的数据，然后清除其中的重复项，清除后的数据有 402,627 条，根据未重复项的项目名称来统计该项目名称出现的次数，得到具体如下所示：

由表 1 职业类型可知，职业类型大体可分为七大类，其中技术类职位需求量最大，占了总体数据的 40.7%，其次为市场与销售，占 20.6，%后面的顺序依次为：运营、设计、职能、产品和金融，所占比例较小。

业由表 2 行业领域看出，在所有行业中，移动互联网占据社会的主导地位，排在第二的是电子商务，第三是金融。由此可看出，在当今这个大环境下，不管是什么行业领域，均以互联网为支撑发展。

由表 3 可以清晰地看到职业前五名，职业词频比例最大的是移动互联网，后面依次为移动互联网 • 电

Table 1. Professional types
表 1. 职业类型

职位	职位词频	职位	职位词频
技术	163,915	运营	67,501
职能	26,840	设计	28,859
市场与销售	83,337	金融	7009
产品	25,166		

Table 2. Industry field
表 2. 行业领域

行业	行业词频	行业	行业词频	行业	行业词频
移动互联网	129,307	硬件	4492	其他	15,098
电子商务	81,951	教育	12,545	广告营销	1362
金融	61,322	旅游	6570	教育	22,545
游戏	16,687	企业服务	33,599	信息安全	4889

Table 3. Professional types
表 3. 职业类型

职业	职业词频	职业	职业词频	职业	职业词频
移动互联网	75,401	移动互联网·企业服务	15,489	电子商务·O2O	8932
移动互联网·电子商务	42,227	移动互联网·教育	12,740	移动互联网·社交网络	8926
金融	39,410	移动互联网·游戏	12,366	移动互联网·硬件	8723
移动互联网·O2O	29,997	移动互联网·文化娱乐	10,301	企业服务	8706
电子商务	29,752	移动互联网·广告营销	9492	O2O	8555
移动互联网·金融	25,434	教育	9317	数据服务	8429
移动互联网·数据服务	17,444	游戏	9216	移动互联网·医疗健康	8276

子商务、金融、移动互联网・O2O 和电子商务等等，从中不难发现，金融、数据服务、企业服务、教育等等，都是在移动互联网的基础上发展，因此应大力发展移动互联网产业，才能推动其它产业的发展。

2.3. 贵州人才需求走向

以贵州招聘信息为研究对象进行具体分析，2016 年 6 月贵州省统计局发布就业报告中显示：今年以来，随着互联网+、新产业、新业态、新商业模式的发展，第一季度组织招聘单位 1760 家，提供就业岗位 10,560 个。网络招聘成为了新的渠道，第一季度我省新注册网上招聘单位 1721 家，与第一季度现场招聘单位数量相当，截至第一季度末，在线招聘单位 14,032 家，提供招聘职位总数 70,160 个，个人注册累计总数 862,900 份，网络招聘已占据半壁江山。

而在实际求职中，如果将求职范围固定在一个区域，那了解本区域人才需求就显得十分重要了。本文以 2015 年 9 月~2016 年 7 月贵阳人力资源网站招聘数据为例，根据贵州省招聘专业的行业、职位等特点，利用目前的人才需求情况，分析对于贵州省的热门行业、职位，并以销售岗位为例展望其未来的人才需求。

由表 4 结合图 1 可知，从供需结构来看，需求较大有销售人员、客服人员、技术支持人员，医疗一系列需求量较小。根据贵阳人力资源网数据显示，2016 年 7 月求职人数前三的专业是通信工程、经济学、测控技术与仪器，求职次数前三的专业分别是通信工程、经济学、电子信息工程。可知 2016 年 7 月通信工程专业需求人员十分饱和。

根据 2016 年 8 月上旬贵州招聘职业类型中类型总数达到 2000 以上描绘条形图，图 2 中可看出在贵州地区，人才需求最大的是行政类岗位，其次是销售类、财务类等、百货类、房地产类等。

通过贵阳人力资源网 2016 年 8 月上旬按国民经济类型分，可看出贵阳热门行业分别为批发和零售业、建筑业以及信息传输、计算机服务和软件业(图 3)。网上招聘单位总数分别约为 200、155、140，这些行业的需求量较大。在 8 月《贵州省大数据产业统计报表制度(试行)》近日获国家统计局批复实施。这是国家统计局批准实施的首个省级大数据产业统计报表制度，标志着贵州省将在全国率先开展大数据产业统计。贵州省内涉及大数据及相关产业的人才需求会渐渐增加，特别是从事大数据研究、规划、管理、应用和实施大数据及相关人员需求，以及通信、互联网等基础设施建设人员。

Table 4. Professional types
表 4. 职业类型

职业类型	总数
销售/客服/技术支持	15,943
生产/营运/采购/物流	4844
服务业	4596
建筑/房地产	4569
人事/行政/高级管理	4380
会计/金融/银行/保险	3089
咨询/法律/教育/科研	2240
计算机/互联网/通信/电子	1817
广告/市场/媒体/艺术	1724
公务员/其他	1043
生物/制药/医疗/护理	576

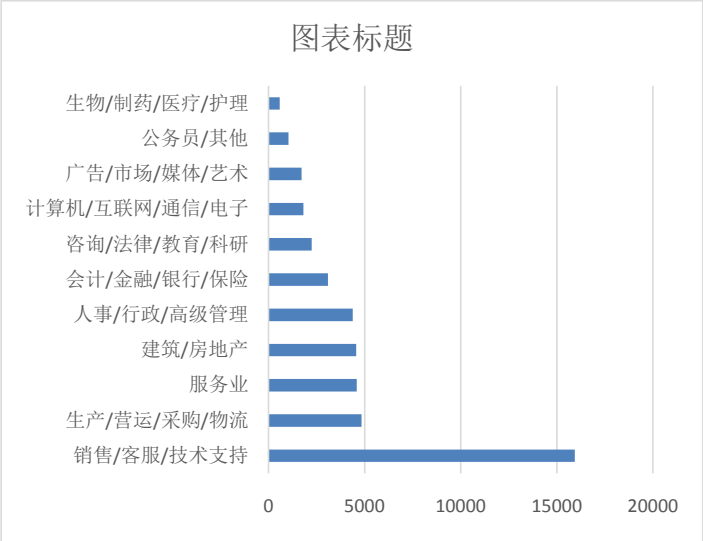

Figure 1. Professional types
图 1. 职业类型

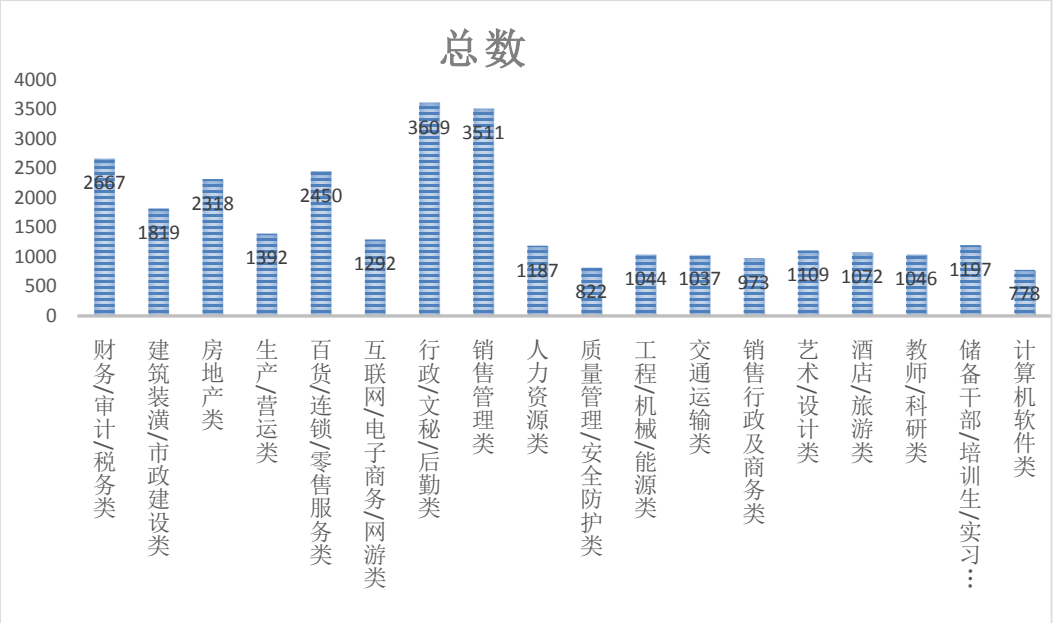

Figure 2. Guizhou province recruitment professional type pie chart
图 2. 贵州省招聘职业类型

将 2015 年 9 月~2016 年 7 月的贵州省销售人员需求量时序列用 x 表示，时序图[3]如图 4 所示，可以看出，该时序图蕴含长期趋势，有一定增长趋势。说明该序列是非平稳时间序列，需要进行平稳化处理。

为了使数据平稳，对它进行差分运算，消除趋势性。故对社会保障支出对数据进行 1 阶差分后，时序图依然含有增长趋势，为非平稳序列。再对其进行 2 阶差分，也为非平稳时间序列，进而再进行 3 阶差分，时序图如图 5 所示，可看出该序列为平稳序列，对其进行白噪声检验，该序列为非纯随机序列。为避免出现过度差分的现象，造成太多信息的损失，而且时序图平稳，所以只进行了 3 阶差分。

图 6 和图 7 的自相关图和偏自相关图可知，除 1 阶偏自相关系数大于两倍标准外，其他的都在两倍

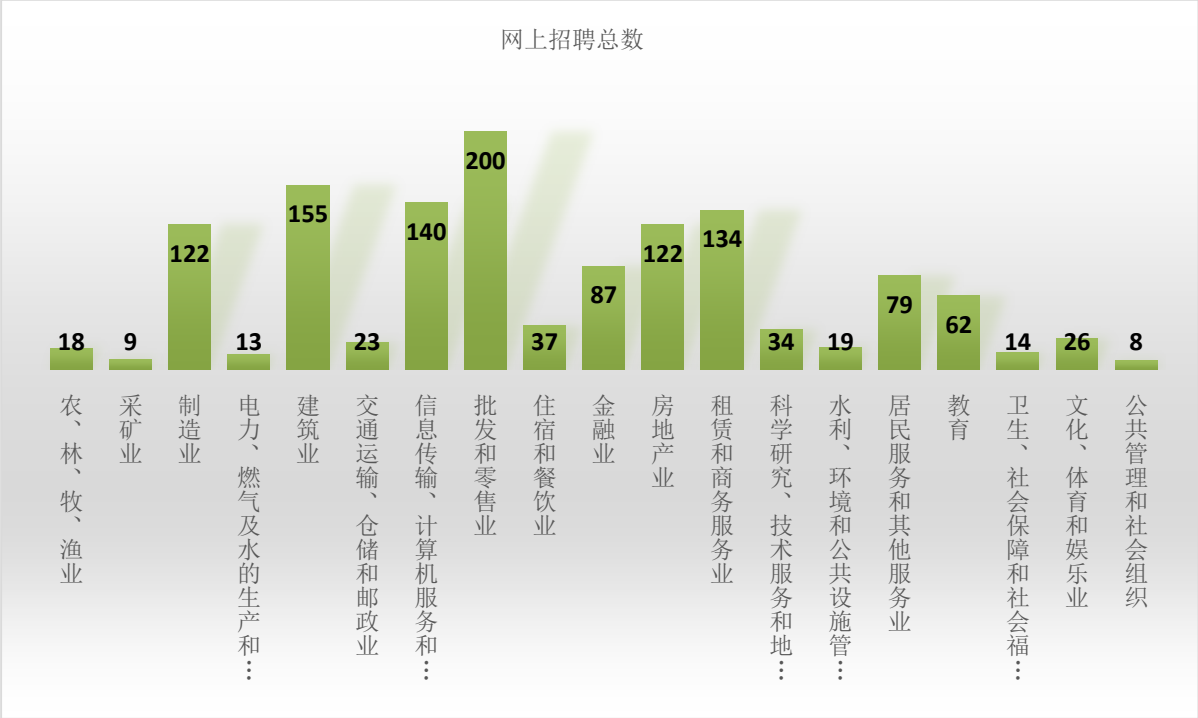

Figure 3. Hiring distribution network
图 3. 网络招聘人数分布图

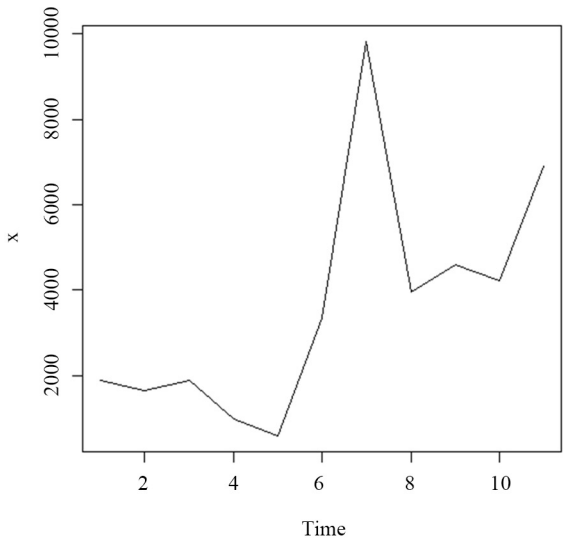

Figure 4. Guizhou province sales staff demand sequence diagram
图 4. 贵州省销售人员需求量时序图

标准差外，二者都是拖尾。自相关系数和偏自相关系数均显示不出截尾的性质，因此可以尝试使用 ARIMA(1,1,0)。经过反复尝试及拟合，得到模型 ARIMA(1,3,1)的 AIC 函数到达最小的模型被认为是最优模型。

确定拟合模型的口径之后，还要对该拟合模型进行必要的检验。模型的显著性检验[4]主要是模型的有效性，显著有效主要看它提取的信息是否充分。换言之，ARIMA 模型(1,3,1)拟合残差项中将不再蕴涵

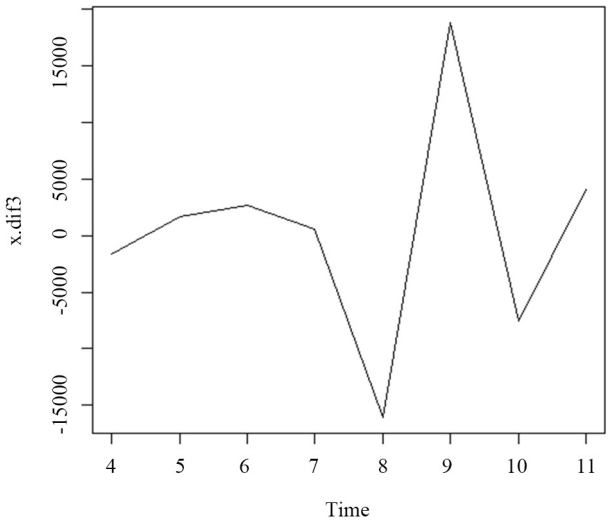

Figure 5. Guizhou province sales staff demand for third-order differential sequence diagram
图 5. 贵州省销售人员需求量三阶差分时序图

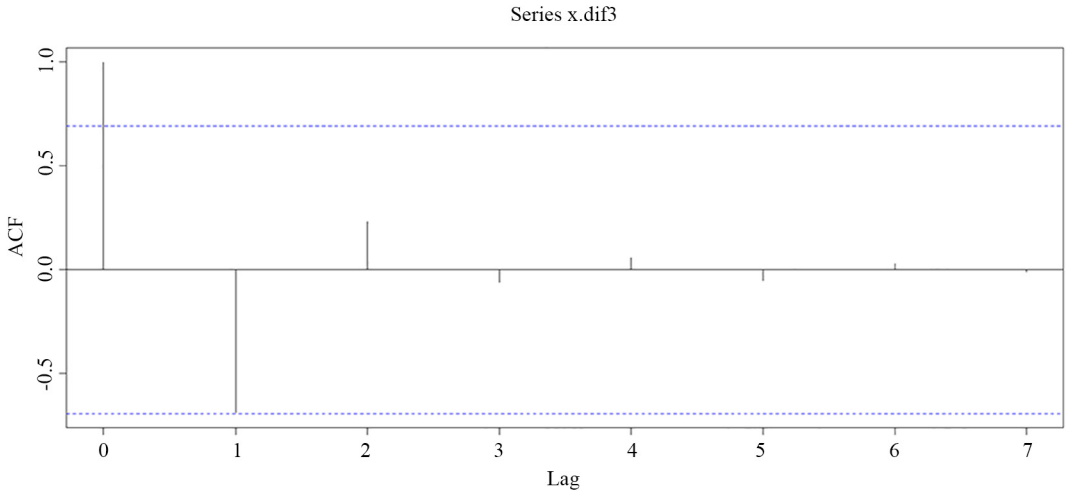

Figure 6. Guizhou province salesmen autocorrelation figure
图 6. 贵州省销售人员自相关图

任何相关信息，即通过白噪声序列检验，残差序列为白噪声序列，该模型称为显著有效模型。模型方程为：

$$(1-B)(1-B^3)x_t = \frac{1}{1+0.4204B-1.9944B^3}\varepsilon_t, \varepsilon_t \sim N(0, 0.0925)$$

根据模型对 2016 年 8 月~2016 年 10 月贵州省销售人员需求进行预测，预测值如表 5。

由表 5 可知，未来三期贵州省销售人员需求人员处在一个不断上升的趋势。

2.4. 大数据相关职位的预测

对数据分析的相关职业进行需求分析，并使用灰色预测法，预测未来数据行业的发展趋势。下面均是用 R 软件画出来的词云图，这些词汇都是招聘单位所给的地域名称、职位名称以及行业名称，首先安装并加载 R [5]中工具包(wordcloud)，利用 SPSS 读取整理好的词频文件，设置好相应的颜色即可，为了

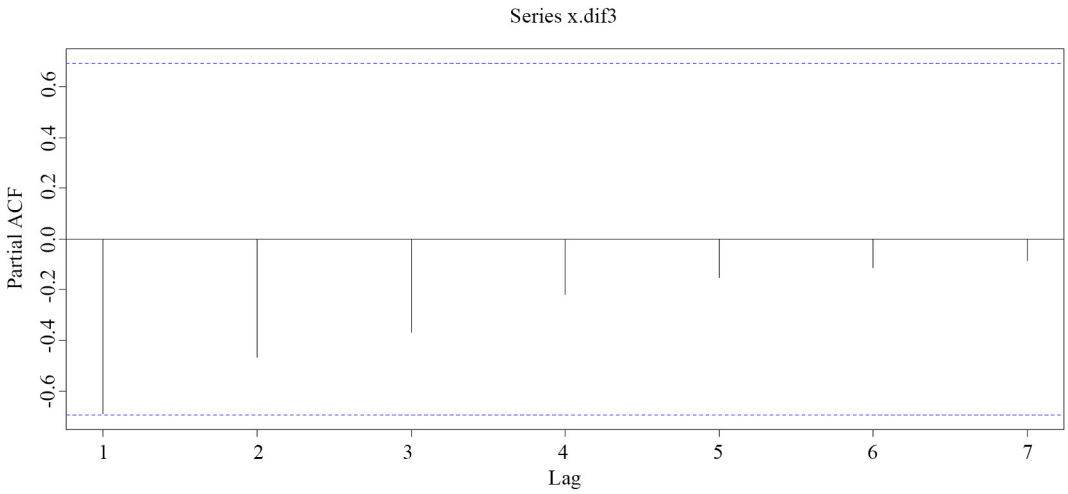

Figure 7. Sales staff in Guizhou partial autocorrelation figure
图 7. 贵州省销售人员偏自相关图

Table 5. Guizhou province sales staff demand forecast
表 5. 贵州省销售人员需求预测值

时间	Point Forecast	Lo 80	Hi 80	Lo 95	Hi 95
2016 年 8 月	8516.014	2397.9922	14,634.04	−840.693	17,872.72
2016 年 9 月	10,922.852	−901.7431	2747.45	−7161.305	29,007.01
2016 年 10 月	13,288.856	−6339.9862	32,917.7	−16,730.868	43,308.58

使统计出来的数据更精准，我们在统计时首先剔除掉重复数据，其次去掉空格的数据，然后再进一步进行词频汇总，接着就画出了词云图如下。

图 8 表示的是地域词云图，其中最明显的城市是北京，其次是上海，深圳、广州和杭州，从地理位置上看它们均属于我国的沿海地区，是我国的“一线城市”，经济发展迅速，对人才的需求量相当大。

紧接着是分析出了关于职业的词云图(图 9)，其中最显眼的是开发，因此在所有的职业中，开发的人数量最大，其次是销售、运营、市场和营销，产品等职业。两个词云图结合了对地域、职业和行业的分析，结果表明，地域毋庸置疑北京的需求职位最多；在职业方面，市场与销售的需求量较大。

针对高级数据工程师、大数据分析师、高级数据开发工程师、数据挖掘工程师等在招聘网站出现的频率，分别间隔 30 天作一次统计汇总，分析与预测相关职位的需求情况，结果如表 6 所示，从表中我们了解到数据挖掘工程师招聘最多，数据库工程师、数据库开发工师二者之间相差不多，数据挖掘与分析高级工程师相对招聘较少。仅仅从数据中进行预测并不科学，因为每一个月招聘数目并不相同，因此，随机抽取几个月数据工程师招聘的信息量分别把数据职位占整个招聘比例计算出来，利用灰色预测法对数据工程招聘比例进行预测，得到的结果如表 7 所示。

我们根据灰色预测得到的预测方程为：

$$x1(t+1) = 1.8394627 \quad e(0.1360886) \quad -1.8394627$$

最后预测下一个月的数值为 0.4137721，因为相对精度等于 1，说明这个模型的精度非常高，具有一定可靠性。从预测数值中我们可以大胆预测数据工程师未来需求将会提高，并提高所占整个招聘行业的比重。

Figure 8. Regional word cloud
图 8. 地域词云图

Figure 9. Professional word cloud
图 9. 职业词云图

Table 6. Frequency table data related industry
表 6. 数据相关行业频率表

职位	Fre	职位	Fre	职位	Fre
数据挖掘工程师	468	数据挖掘工程师	468	数据挖掘工程师	500
数据库工程师	309	数据库工程师	318	数据库工程师	379
数据库开发工程师	221	数据库开发工程师	216	数据库开发工程师	288
数据库管理员	173	数据库管理员	146	数据库管理员	145
数据仓库工程师	101	数据仓库工程师	99	数据仓库工程师	121
大数据架构师	93	大数据架构师	67	大数据架构师	97
高级数据开发工程师	37	高级数据开发工程师	45	高级数据开发工程师	33
高级数据工程师	21	高级数据工程师	31	高级数据工程师	29
大数据数据分析师	2	大数据数据分析师	5	大数据数据分析师	3
数据挖掘与分析高级工程师	1	数据挖掘与分析高级工程师	2	数据挖掘与分析高级工程师	1

Table 7. Hiring proportion
表 7. 招聘比例

月份	1	2	3
整个招聘比例	0.3225	0.3157	0.3618

3. 结论

通过以上的分析概括，我们了解了当今社会的热门职业类型和行业领域都在哪些方面，为了便于实际的研究分析，我们还对贵州省销售人员需求作了分析预测，模型自身的先天性缺陷在于随着预测期的延长，其预测误差会逐渐增大，但与其他的预测方法相比，在短期内其预测的准确程度较高。根据以上的对比，发现 ARIMA 模型能较好地分析、计算贵州省销售人员需求的发展波动的情况，具有较好的应用前景，数据来源于网络，可能存在误差，使用灰色预测法只能用对应关系的数据进行预测，不具有代表性。因此本文采用招聘比例进行预测，使代表精度远远提高，更能了解社会和相关行业的需求特点以及发展趋势。

基金项目

贵州师范学院校级学生科研项目(项目编号：2016DXS097)；贵州省 2014 年省级本科教学工程项目“计算机科学与技术”专业综合改革(项目编号：黔教高发[2014]378 号)；卓越工程师教育培养计划项目(黔教高发[2013]446 号)；2015 年省级本科教学工程建设项目(黔教高发[2015]337 号)；2016 年大数据视角下的贵阳市交通优化配置问题研究(项目编号：201614223037)。

参考文献 (References)

[1] 王宇昕. 大学生动态求职招聘与信息分析系统的分析与设计[D]: [硕士学位论文]. 北京: 北京邮电大学, 2012.
[2] 钟晓旭. 基于 Web 招聘信息的文本挖掘系统研究[D]: [硕士学位论文]. 合肥: 合肥工业大学, 2010.
[3] 贾俊平. 统计学——基于 R [M]. 北京: 中国人民大学出版社.
[4] 茆诗松, 吕晓玲. 数理统计学[M]. 北京: 中国人民大学出版社.
[5] 王燕. 时间序列分析——基于 R [M]. 北京: 中国人民大学出版社.